STAT 651 Chapter 3.1–3.6.1

Zepu Zhang
December 16, 2010

Sections 3.1–3.3 discuss the more common univariate discrete and continuous distributions. Although you should read it through, you won't be able to remember everything here.

The following distributions are mentioned:

Discrete: discrete uniform, Bernoulli, hypergeometric, binomial, Poisson, negative binomial, geometric

Continuous: uniform, gamma, chi squared, exponential, Weibull, normal, beta, Cauchy, lognormal, double exponential

Most of these are very "basic" distributions—variations and transformations of these can create other distributions. Important distributions that are in a sense "derived" or "for comparison" are not discussed here, e.g. student t distribution and F distribution.

Of these, probably the more important ones (for general or our purpose) are binomial, Poisson, uniform, gamma, exponential, beta, and normal. An indication of their importance is that you will come to memorize some formulas (pdf's) of these distributions, perhaps except gamma and beta, whereas for the others you do not need to make the attempt—it suffices to know where to look it up. Discrete and continuous uniform are trivial but useful; geometric is easy to remember.

It is very important to come to appreciate that different distributions are used to model different types of phenomena. (Indeed they arose as abstractions of different types of typical problems.)

Speaking of application, the first thing to check is the range of the value i.e. support of the distribution. The first division here is discrete versus continuous.

Discrete

(a) Finite set of specific values: discrete uniform (or any distribution with given pmf).

(b) Non-negative integers: arising from counting—binomial, geometric, hypergeometric,...

(c) Poisson is kind of special: typical problems of its application are more like "occurrence" than "counting", and the problem has a blend of discrete and continuous setting—e.g. occurrence of earth quakes over time, or appearance of something in space (the count is discrete, but the “stage” of the occurrence is continuous).

Continuous (a) Whole real line. The grand representative is normal. Others mentioned here are double exponential and Cauchy.

(b) Half real line. For positive variables, there are gamma (including chi squared and exponential), lognormal, and Weibull. Note there is not need to specially invent and study distributions for **negative** variables: they are the opposite of positive variables. Neither is there need to specially deal with, say, $[1, \infty)$ or $[a, \infty)$ for arbitrary a—it's a simple shift from $[0, \infty)$.

(c) Finite-length interval. It suffices to define and study distributions on $[0, 1]$: beta, uniform. Any other interval is a simple shift and scaling of this.

Some distributions have pretty clear application targets:

1. Bernoulli, binomial, geometric, hypergeometric, negative binomial all correspond to particular experiment settings or counting problems.

2. Poisson is widely used for counting the occurrence of something in space or time that has an fixed rate of occurrence (i.e. “intensity”); the occurrence during different periods or in different areas is independent.

3. Both negative binomial and Poisson may be used to model a phenomenon in which we wait for some occurrence; see example 3.2.6 on p. 96. Geometric and exponential have similar uses if their “memoryless” property is desired.

4. The wide applicability of normal is to a large extent guaranteed by the Central Limit Theorem. For example, random error (from measurement or model) is usually modeled by normal. Normal is also a good approximation for some things in asymptotic (large sample) settings.

5. The “lack of memory” property of geometric and exponential makes them useful for modeling certain types of “life time”, “waiting time”, etc.

6. Weibull can be used to model failure time data; alluded to on p. 102.

7. Cauchy could be used to model the ratio of two random variables.

8. Beta is a good tool for modeling proportions, thanks to its support $[0, 1]$.

9. Lognormal is used in modeling many positive, right skewed quantities in science and engineering. Its relation with normal makes it easy to use mathematically. (The log of the positive quantity is normal.)

The book shows examples of some properties of a distribution that we typically will pay attention to when considering their applications:

1. Shape, skewness, symmetry: the “shape” and “scale” parameters of gamma (bottom of p. 99); some gamma density curves, figure 3.3.6(b), p. 110; the flexibility of beta, p. 107; figure 3.3.3, p. 107 and figure 3.3.4, p. 108; the right-skewness of lognormal, figure 3.3.6(a), p. 110.

2. Thickness of tail: Normal has particularly thin tails. This suggests that extreme values are very hard to happen in a normal distribution. For example, with $> 99\%$ probability a normal variable is with 3σ of its mean (p. 104). This makes normal a bad tool for modeling extreme values. Both Cauchy (p. 108; figure 3.3.5, p. 109) and double exponential (p. 110) have thicker tails than normal. Another such distribution is student t (later).

Distributions are related, e.g.

1. Binomial arises from independent, identical (meaning: same success rate) Bernoulli trials. A Binomial variable is the sum of i.i.d. Bernoulli variables. (P. 89)

2. Geometric is a special case of negative binomial (p. 97).

3. Poisson is a limiting case of negative binomial; p. 96. (I think this fact is less useful.)

4. Exponential and chi squares are special cases of gamma.

5. Exponential is also a special case of Weibull.

6. The Poisson and gamma are related; example 3.3.1, p. 100 (less important).

7. The ratio of two normal variables is Cauchy (p. 108).

8. A chi squared variable of k df is the sum of k independent Z^2, where Z is standard normal.

9. Lognormal is related to normal.

10. Double exponential is the reflection of exponential

Another kind of relation (somewhat not as close as the above, but not necessarily) is approximation. In this age, the approximations are barely important for computation. Their value lie more in theoretical discussions (for example, if, under certain conditions, something is approximately normal, then we can make use of properties of normal distribution in discussion).

1. Approximating binomial by Poisson (p. 93–94).

2. Approximating binomial by normal (p.105).

These approximations are less important to us.

For discrete distributions, we focus on their pmf. For continuous distributions, we focus on their pdf. For uniform, normal, and exponential, we should be familiar with their cdf as well. (The normal cdf is conceptually important, although it does not have an analytical form.)

These sections introduce (or use) several technicalities that are good to know, including

1. The gamma function and several properties of it; see (3.3.2), (3.3.3), (3.3.4), (3.3.15).

2. Integral of the "normal kernel"; see (3.3.14) on p. 103.

3. Changing integration from Cartesian to polar coordinates; see bottom of p. 103.

4. The beta function and its relation to the gamma function; p. 106.

5. The binomial theorem: 3.2.2 on p. 90.

6. The Taylor series expansion of e^x, middle of p. 92.

7. Sum of converging geometric series, lower part of p. 92.

8. Dealing with integrals containing absolute values; see p. 110–111.

1 Bernoulli

pmf:

$$f(x) = \begin{cases} p, & x = 1; \\ 1 - p, & x = 0 \end{cases}$$

Mean:

$$E(X) = 1 \cdot p + 0 \cdot (1 - p) = p$$

$$E(X^2) = 1^2 \cdot p + 0^2 \cdot (1 - p) = p$$

Variance:

$$\text{var}(X) = E(X^2) - \big(E(X)\big)^2 = p - p^2 = p(1 - p)$$

If we define $q = 1 - p$, then $\text{var}(X) = pq$.

2 Binomial

Two ways to define (and understand) binomial:

1. Number of “success” in n independent, constant success rate, trials.
2. Sum of n i.i.d. Bernoulli variables.

Some use the notation $\mathbf{Bin}(n, p)$, or $\text{Binomial}(n, p)$.

pmf:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, \ldots, n$$

mean (example 2.2.3, p. 56): np.
(The binomial mean can be understood intuitively.)

variance (example 2.3.5, p. 61): $np(1 - p)$.

mgf (example 2.3.9, p. 64): $\big[pe^t + (1 - p)\big]^n$.

Via the 2nd definition (as sum of i.i.d. Bernoulli), the binomial mean and variance are apparent.

cdf: may calculate using computer, or by normal approximation (p. 105).

Theorem 3.2.2, Binomial theorem (p.90).

Exercise Show the binomial probabilities sum to 1.

Example 3.2.3, p. 91.

3 Poisson

pmf:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, \ldots$$

"intensity" λ.

Exercise Show the Poisson probabilities sum to 1.

mean: λ

variance: λ

mgf: two ways: (1) as we did before; (2) using Taylor series expansion of e^x.

Typical situation of application: Consider some point event occurs as time goes by. Assumptions: (1) The probability of occurrence is λ per unit duration. This behavior does not change with time. (2) Occurrence in non-overlapping time intervals are independent of each other. Then the probability of k occurrence in a time interval of length t is $P(X = k \mid \lambda t)$. This stochastic process is called a Poisson process. Analogously in space; just replace "duration" by "area".

Example 3.2.4, p. 93.

4 Geometric

Know the meaning of geometric, and based on that understand its pmf:

$$P(X = x) = (1 - p)^{x-1} p, \quad x = 1, 2, \ldots$$

Exercise See how geometric is generalized to negative binomial, and understand the pmf of negative binomial.

Two ways to derive the cdf of geometric:

(1) $P(X \le x) = p \sum_{k=1}^{x} (1 - p)^{k-1} = \ldots$, using the formula for geometric series (p. 31).

(2) $P(X \le x) = 1 - P(X > x) = 1 - (1 - p)^x$

Example Show the "memoryless" property:

$$P(X > s \mid X > t) = P(X > s - t), \quad s > t$$

Example 3.2.7, page 98.

5 Uniform

pdf, cdf, mean, median, variance

6 Gamma

6.1 The gamma function and its properties

Definition:

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, \mathrm{d}t, \quad x > 0 \text{ (for our purpose)} \qquad (3.3.2)$$

Properties and special values:

$$\begin{aligned} \Gamma(x+1) &= x\Gamma(x) && (3.3.3) \\ \Gamma(1) &= 1 \\ \Gamma(n) &= (n-1)! && (3.3.4) \\ \Gamma\Big(\frac{1}{2}\Big) &= \sqrt{\pi} && (3.3.15) \end{aligned}$$

Exercise Prove these properties.

6.2 The gamma pdf

Now we construct a pdf based on the shape of $x^{\alpha-1}e^{-x}$ but allow a horizontal scaling, that is, replace x by x/β, then the curve is $(x/\beta)^{\alpha-1}e^{-x/\beta}$. To make it a pdf, we need to normalize it by the integral

$$\int_0^\infty (x/\beta)^{\alpha-1} e^{-x/\beta}\, \mathrm{d}x = \beta \int_0^\infty y^{\alpha-1} e^{-y}\, \mathrm{d}y = \beta\Gamma(\alpha)$$

Therefore we define the gamma pdf

$$f(x \,|\, \alpha, \beta) = \frac{1}{\beta\Gamma(\alpha)} (x/\beta)^{\alpha-1} e^{-x/\beta} = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad x, \alpha, \beta > 0$$

α: shape parameter
β: scale parameter

Exercise We have seen β provides a simple horizontal shrinking/stretching of the curve. Let's see how α affects where the curve peaks.

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = 0 \Rightarrow x^{\alpha-2}\big(x - \beta(\alpha-1)\big) = 0$$

Discussion:

(1) $\alpha < 1$: decreasing on $(0, \infty)$.

(2) $\alpha = 1$: exponential, decreasing on $(0, \infty)$.

(3) $1 < \alpha < 2$: $f(0) = 0$; peaks at $\beta(\alpha - 1)$.

(4) $\alpha = 2$: $f(0) = 0$; peaks at β.

(5) $\alpha > 2$: $f(0) = 0$; peaks at $\beta(\alpha - 1)$; flat at 0.

When $\alpha > 1$, the curve gets more and more symmetric as α increases.

This discussion shows the versatility of the gamma distribution.

6.3 Properties

mean

variance

mgf (example 2.3.8 on p. 63)

Note the "recognizing another gamma kernel" technique used in finding the mean, variance ($E(X^2)$), and mgf of gamma.

6.4 Special cases

Exponential: $\alpha = 1$

Chi squared with p df: $\alpha = p/2$**,** $\beta = 2$. Connection with (standard) normal.

7 Exponential

Find the cdf, mean (example 2.2.2, p. 55), variance (example 2.3.3, p. 59), and mgf of exponential.

You should be able to recognize the pdf and cdf of exponential immediately.

The "memory-less" property of exponential (most easily seen from the cdf) and implications for application.

8 Beta

The beta function and its connection with the gamma function.

Definition of the density.

Application: very useful due to its bounded support, $(0,1)$. Used for modeling variables on a bounded support, such as proportions.

Versatility of the curve: p. 107.

9 Normal

9.1 Why is it so important?

The book lists three reasons: (1) analytic tractability; (2) symmetric, bell shape; (3) Central Limit Theorem.

9.2 PDF and CDF

The standard normal pdf.

The integral $\inf_0^\infty e^{-z^2/2}\,\mathrm{d}z$ (pp. 103–104).

The standard vs non-standard normal: usually all computation about a normal variable is delegated to the standard normal.

The standard normal cdf: many use the notation

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,\mathrm{d}t$$

This is no analytical form for this integral. The numerical evaluation of $\Phi(x)$ (the cdf) and $\Phi^{-1}(x)$ (the quantile function) are basic routines that are needed often.

Exercise Find the mgf, mean, and variance of standard and non-standard normal.

9.3 Some useful points

1. Symmetric, bell shape: suitable for modeling many phenomena.

2. Measurement (and model) error is almost always modeled by normal; the validity is provided by the CLT (Central Limit Theorem).

3. The density is very smooth: infinitely differentiable.

4. The distribution is completely specified by two parameters: mean and variance.

5. The symbols μ and σ^2 are very well established. In addition, Z is often used for a standard normal variable.

6. $N(\mu, \sigma^2)$ means the normal distribution, e.g. $X \sim N(3, 10)$.

7. Normal approximation to binomial appears in many intro texts.

8. Normal has famously "thin" tails, making it inappropriate for modeling extreme, rare values, because extreme values occur very rarely in a normal model, making it very inefficient to study them via normal simulation.

9. It is conventional to use standard deviation σ as the unit while measuring how far a value is from the distribution's center. See page 104, middle. "Three sigma" (from the center) typically suggests a pretty extreme value.

10. A nice companion to normal is the t distribution, which
(1) has a similar nice bell shape;
(2) has thicker tails than normal;
(3) approaches normal as its df increases, hence its closeness to normal is adjustable by a single parameter (the degree of freedom).

11. Multivariate normal has additional nice properties, to be studied later.

10 Lognormal

Exercise Derive the pdf.

11 Double exponential

I want to use the derivation of the mean and variance of double exponential to demonstrate two techniques.

(1) Partition the support so that the absolute value sign is stripped, and integration can proceed. See pages 110–111.

(2) More intuitive but equally rigorous thinking can go like this:

Mean Note the pdf is symmetric about μ, therefore the mean must be μ if the mean exists.

Variance First, we can shift the distribution to be centered at 0 without affecting the variance. Second, how is the symmetry useful here? Now the variance is $E(X^2)$. Because of the symmetry, I would say the variance is equal to the $E(X^2)$ on either half of the support. With this observation, we can get the variance through that of an exponential distribution.

12 Exponential families

Definition (3.4.1), page 111. A pdf or pmf belongs to the "exponential family" if it can be expressed as

$$f(x \,|\, \vec{\theta}) = h(x)\, c(\vec{\theta})\, \exp\Big(\sum_{i=1}^{k} w_i(\vec{\theta})\, t_i(x)\Big)$$

where

1. $h(x) \geq 0$, $c(\vec{\theta}) \geq 0$ (hence $f(x) > 0$).
2. $h(x)$ and $t_i(x)$ do not depend on $\vec{\theta}$, and $c(\vec{\theta})$ and $w_i(\vec{\theta})$ do not involve x.
3. x here is univariate but the parameter $\vec{\theta}$ can be a vector (i.e., have multiple components).
4. $h(x)$ may incorporate an indicator function that shows the range of x. See definition 3.4.5, page 113. The indicator function must not involve $\vec{\theta}$.

Example 3.4.1, page 111. Binomial. (n is not considered a parameter, i.e. an element of $\vec{\theta}$, here.)

Example 3.4.4, page 113. Normal.

Example Show Bernoulli, $f(x \,|\, p) = p^x(1-p)^{1-x}, x = 0, 1$, belong to the exponential family.

Example Show Poisson, $f(x \,|\, \lambda)$, belongs to the exponential family.

Example Show gamma, $f(x \,|\, \alpha, \beta)$, belongs to the exponential family.

Exercise Show geometric, $f(x \,|\, p)$, belongs to the exponential family.

Example The uniform, $f(x \,|\, \theta) = \frac{1}{\theta} I_{(0,\theta)}(x)$, is not a member of the exponential family. There is no way to separate θ out of the indicator function.

Theorem 3.4.2, and example 3.4.3, page 112. Self reading as calculus exercise.

Exercise Find the variance of binomial following example 3.4.3 and using theorem 3.4.2.

Section 3.4 from (3.4.7), p. 114, onwards can be skipped (unless we find it needed later).

13 Location and scale families

Given a continuous density $f(x)$, it is easy to create a whole family of related distributions by a simple transformation: let $Z = \sigma X + \mu$, then $f_Z(z) = \frac{1}{\sigma} f_X\big(\frac{z-\mu}{\sigma}\big)$ is a valid pdf.

The Z pdf has the same “shape” as the X pdf, only shifted horizontally by the amount of $|\mu|$ and scaled (compressed or stretched) horizontally by σ or $1/\sigma$.

μ: location parameter; σ: scale parameter.

Example The normal distribution,

$$f(x \,|\, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\biggl(-\frac{(x-\mu)^2}{2\sigma^2}\biggr)$$

is a location-scale family. In this family, the “standard member” is the standard normal $f(x \,|\, \mu = 0, \sigma = 1)$. Denote a standard normal variable by Z, then $X = \sigma Z + \mu$ is normal with mean $E(X) = \sigma E(X) + \mu = \mu$ and variance $\operatorname{var}(X) = \sigma^2 \operatorname{var}(Z) = \sigma^2$.

Remark 1. The idea of location family can be used to accommodate bounds of the variable. For example, exponential is defined on $(0, \infty)$. If we want to model a random variable that is > 100 but like to use the exponential shape, we can shift the exponential to the right by 100, that is, define $X = Z + 100$ where Z is exponential. (Of course, a shift to the left is equally fine.)

2. The “standard” member is usually chosen to be the one with mean 0 and variance 1 (if possible), e.g. standard normal.

3. If location and scale are both applied, think scaling, then shifting (for visualizing the pdf curve). For example, $\frac{1}{8}e^{-(x-3)/8}$ is the exponential e^{-x} stretched 8 times horizontally, then shifted to the right by 3.

4. Calculation usually uses the standard member of the family as a reference, because both pdf and cdf of X are straightforwardly connected with those of Z. For example, in any statistical software, it must be the case that the pdf and cdf

of normal distributions are implemented for the standard normal only.

14 Chebychev's Inequality

Theorem 3.6.1, page 122.

Note the condition: $g(x) \geq 0$ (and then naturally $r > 0$).

Proof.

Example 3.6.2, page 122.

Example 3.6.3, page 123.

Example 3.45(a), page 134.